УДК 519.7

Н.В. Борисова, О.В. Канищева, Е.Н. Юрченко

АВТОМАТИЗИРОВАННОЕ ФОРМИРОВАНИЕ СЛОВАРЯ СИНОНИМОВ

Введение. Для решения многих задач автоматизированной обработки естественного языка часто необходимо иметь словарь синонимических терминов. Электронных словарей синонимических терминов существует не так много. Отметим, что построение компьютерных словарей, тезаурусов и грамматик – объемная и трудоемкая работа, иногда даже более трудоемкая, чем разработка лингвистической модели и соответствующего процессора. Поэтому одной из подчиненных задач компьютерной лексикографии является автоматизация построения лингвистических ресурсов.

Компьютерные словари часто формируются конвертацией обычных текстовых словарей, однако нередко для их построения требуется гораздо более сложная и кропотливая работа. Обычно это бывает при построении словарей и тезаурусов для быстро развивающихся научных областей – молекулярной биологии, информатики и др. Исходным материалом для извлечения необходимой лингвистической информации могут быть коллекции и корпусы текстов.

Цель работы. Авторы данной статьи видят свою задачу в лингвистической формализации критериев синонимичности с целью создания автоматизированного синонимического словаря, принцип работы которого будет основан на построении синонимических рядов из слов, описанных в толковых словарях. Программно реализовать предложенную формальную модель.

Основной материал. Рассмотрим определение термина "синоним". До сих пор причиной большинства разногласий у исследователей синонимии было определение степени "близости" или "тождества" значений слов-синонимов. Однако для современных ученых этот вопрос был решен еще в 60-е гг. XX века. "Наиболее распространенным пониманием синонима является следующее: ***синонимами*** признаются слова, выражающие одно и то же понятие, ***тождественные*** или ***близкие*** по своему значению, которые отличаются один от другого или оттенками значения, или стилистической окраской (и сферой употребления), или одновременно обоими названными признаками" [1].

В подтверждение данной точки зрения Л. А. Чешко [1] приводит ссылки на работы Ю. Д. Апресяна, А. Н. Гвоздева, А. Б. Шапиро и А. П. Евгеньевой. Та же точка зрения к определению синонимов прослеживается и у современных украинских ученых: "В основу Словника покладено розуміння синонімів як ***одиниць*** з ***тотожною*** або максимально ***наближеною*** *предметно-поняттєвою віднесеністю*, що виступають у значенні тієї самої частини мови" [3]; "Синонімія − …наявність ***однакового*** чи ***близького*** *змісту* у кількох ***знаків*** одного статусу (рівневого, частиномовного тощо)" [4].

Цитируемые определения отражают традиционный взгляд на проблему и в свете приближения к решению поставленной задачи лишь акцентируют внимание на том, насколько тесно проблема синонимии связана с "предметно-понятийной отнесенностью", толкованием или значением слова.

Обратившись к "Предисловию" к Новому объяснительному словарю синонимов русского языка, созданному под общим руководством академика Ю. Д. Апресяна и изданному в 2003 г. [5], находим наиболее современное, на наш взгляд, определение: "Синонимы − лексемы, ***толкования*** которых доведенные до уровня семантических примитивов, ***имеют такую пересекающуюся часть***, ***которая либо больше их суммарных значений*** (если рассматриваются две лексемы), ***либо не меньше, чем их суммарные различия*** (если рассматриваются три и более лексем)" [5], которое хотя и подтверждает наши интуитивные предположения о том, что решение задачи находится в области семантики, но практически её решить не помогает, поскольку, по мнению Ю. Д. Апресяна, приведенное выше определение "…не может считаться вполне формальным, потому что разные семантические примитивы могут иметь разный вес в семантической системе данного языка". Из выше сказанного следует, что слова-синонимы отдельно взятого синонимического ряда будут отличаться оттенками значения, "периферией", а близкими или тождественными у них должно быть понятие, обозначаемое разными терминами: "ядро значения", "семантический примитив", "ядерная сема" или "компонент-доминанта" [6].

На рисунке 1 мы видим, что лексическое значение слова, представленное в виде "поля" [6], помогает понять сущность слов-синонимов, значения которых, совпадая по одним семемам, разняться по другим. Лексическое значение слова фиксирует разные аспекты мысленного отображения предметов и явлений: денотативный аспект связан с представлением признаков и свойств обозначаемых предметов, сигнификативный аспект представляет его понятийный смысл. Денотативный и сигнификативный аспекты создают предметно-понятийное ядро лексического значения (интенсионал). Это самый стойкий компонент значения, которому противопоставляется изменяемая смысловая периферия (импликационал). Это целый комплекс предметно-образных связей и отношений, где объективируется не четко детерминированная (обусловленная), а вероятностная структура обозначаемого фрагмента действительности. Эволюция лексических значений слов возможна благодаря этим свойствам языка [2].

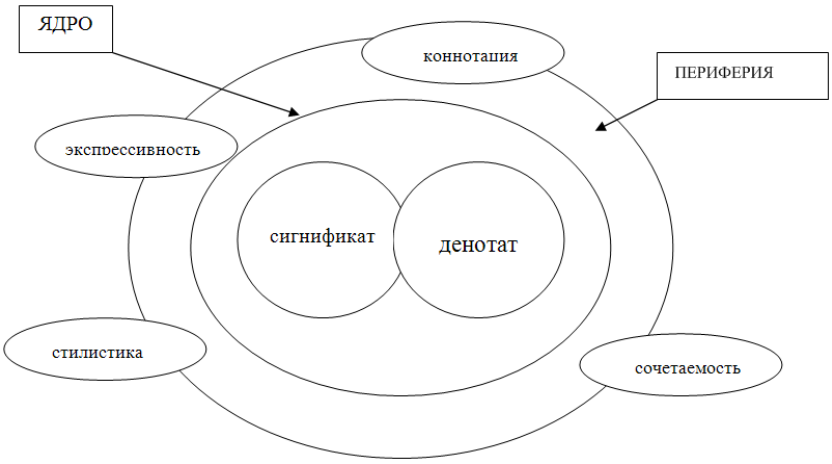

Рисунок 1 – Семантическая структура слова

Представив схематично традиционное "линейное" восприятие синонимического ряда (рис. 2) в виде синонимического поля (рис. 3), становится понятным, чем именно отличаются слова-синонимы, и какой принцип положен в основу существующих классификаций синонимов.

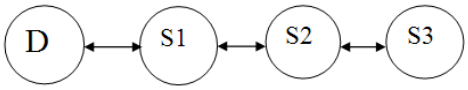

Рисунок 2 – Синонимический ряд

Синонимический ряд, изображенный на рис. 2 линейно, представляет собой доминанту (D) и несколько синонимов (S1, S2, S3), которые, как видно из отношений, обозначенных стрелками, синонимичны доминанте и друг другу.

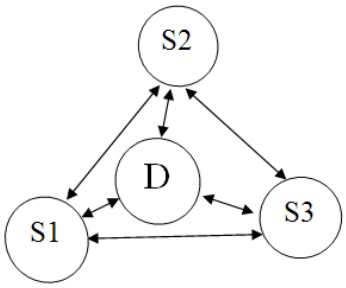

Рисунок 3 – Синонимическое поле

Представление синонимических отношений в виде "поля" является, на наш взгляд, более наглядным в плане классификации синонимов и более точно соответствует определению синонимического ряда, в котором все синонимы должны быть синонимичны доминанте и друг другу.

Таким образом, следуя схеме семантической структуры слова (рис. 1), мы приходим к выводу, что значения "близких по значению слов" − синонимов различаются периферийными семами: экспрессивностью, коннотацией, стилистическими признаками и сочетаемостью. Существующие классификации в обобщенном виде выделяют следующие типы синонимов [4]: ***лексические*** и ***контекстуальные*** − первые имеют сходство в языковой системе, вторые получают ее в контексте. В данной работе нас интересуют лексические синонимы, поскольку именно их поиск планируется вести по толковому словарю. Среди них выделяют ***абсолютные*** (например, "алфавит" и "азбука", имеющие практически идентичные толкования: ср. "Алфавит − установленный порядок букв, используемых в системе письма, характерного для определенного языка" и "Азбука − совокупность букв, используемых в системе письма, построенного на основе славянских букв; алфавит" [7]) и ***частичные***, т.е. "тождественные" и "близкие по значению", которые, в свою очередь, подразделяются на:

- *семантические* (идеографические, ближняя периферия) − близкие по значению, но не тождественные по семному составу слова, которые передают разные оттенки значения одного понятия и могут определяться в толковом словаре одно через другое (напр.: "Дом − жилое здание, строение" − "Здание − строение, сооружение, постройка (обычно больших размеров)" − "Строение − здание, постройка" [7]);
- *стилистические* (функциональные) − отличаются сферой употребления, их периферийные семы наиболее очевидны; формально в толковых и синонимических словарях обычно отличаются

пометами, которые традиционно даются в сокращенном виде, чаще курсивом (напр. дом − *оф.* жилище [7] − *разг.* домина − *уст.*, *обл.* хоромина − *уст.* храмина [5]).

- *коннотативные* (эмоционально-оценочные) − отличаются отношением говорящего к названому им явлению; формально также отличаются пометами (напр. "Дом − *уст.*, *шутл.* хоромы");
- *семантико-стилистические*, которые совмещают вышеперечисленные признаки и отличаются семантическим значением, эмоциональной окраской и сферой употребления.

Так, например, в Электронном синонимическом словаре-справочнике В. Н. Тришина [8] синонимов к слову "дом" насчитывается более 100, см. рис. 4.

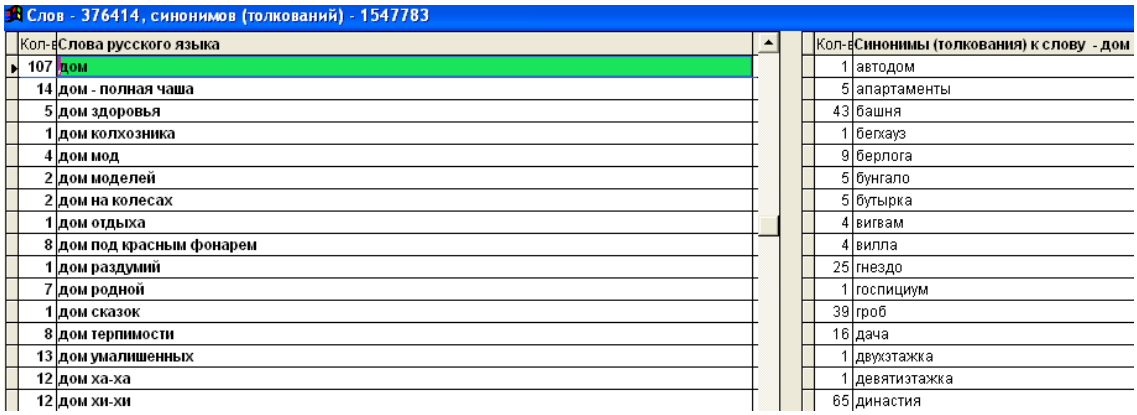

Рисунок 4 – Электронный синонимический словарь-справочник В. Н. Тришина

Что касается сочетаемости − одна из периферийных потенций значения (рис. 1), − то в зависимости от нее выделяются так же:

- **синтаксические синонимы** – грамматически разные конструкции, выражающие одну и ту же мысль;
- **морфологические синонимы** − варианты словоформ, передающих то же понятие (напр., родительный падеж: чаю − чая, меду − меда и т.п.).
- **фразеологические синонимы** – разные фразеологизмы с одним значением (напр., ни туда ни сюда − ни в тын ни в ворота).

В данном исследовании было решено ограничиться однословными синонимами. Резюмируя вышесказанное, мы приходим к определению ***синонимического ряда***, который представляет собой семантико-синонимическое поле слов тождественных или близких по значению, ядром которого является доминанта − наиболее употребительное, стилистически нейтральное, наименее экспрессивное и синтагматически наименее фиксированное слово [4].

Рассмотрим данное определение на примере слова "голова", к которому будем строить синонимический ряд, используя толковый словарь ABBY Lingvo 12: "**ГОЛОВА** I 1. Верхняя часть тела человека, передняя или верхняя часть тела позвоночного животного, состоящая из черепной коробки и лица у человека или морды у животного. Волосы на этой части тела человека (обычно о прическе). 2. *перен.* Ум, рассудок, сознание. 3. *перен.* Передняя часть движущейся группы, колонны. II *разг.* Животное или птица как единица счета". Берем только первое значение слова "голова". Поскольку это значение наиболее нейтральное (нет помет), то принимаем слово "голова" за доминанту. Толковый словарь, из которого мы взяли данное определение слова "голова" дает ссылку только на один синоним: "Голова Syn: глава". Проверяем его по тому же словарю: "**ГЛАВА** – *устар.* То же, что **голов**а". Однако, исходя из данных электронного синонимического словаря-справочника В. Н. Тришина, синонимический ряд к слову "голова" насчитывает более 100 слов-синонимов. Чтобы построить синонимический ряд по толковым словарям хоть в какой-то мере количественно приближенный к ряду, приведенному в словаре-справочнике В. Н. Тришина, слово "голова" принимается за доминанту, а его основа за "ядерную сему" значения (выделено в примерах полужирным), которая должна совпадать в значениях слов-синонимов:

БАШКА – *разг.-сниж.* **Голов**а как часть тела человека или животного

БАШНЯ – *шутл.* О **голов**е человека.

КОТЕЛОК – *разг.* **Голов**а, умственные способности.

ЧЕРДАК – *разг.-сниж.* О **голов**е, уме.

ГОРШОК – *разг.* О ровной линии вокруг **голов**ы, делая волосы одной длины.

ЧЕРЕП – скелет **голов**ы и позвоночных животных. *Разг.* Верхняя часть **голов**ы. *Разг.* **Голов**а без волос.

РЕПА – *разг.* О глупой или плохо соображающей **голов**е.

Таким образом, наша идея заключается в том, чтобы из массива сло́вника толкового словаря (в электронной форме) автоматически вычленить синонимические ряды, следуя двум принципам (из трех)

построения синонимического ряда, которые нам удалось формализовать:

1. ***Семантическому***, т.е. поиск ведется по "ядерной семе" доминанты (морфологически подготовленной основе или корню доминанты) в части толкований значений слов.

Этот принцип должен работать в любом электронном толковом словаре, но он противоречит "полевому принципу" построения синонимического ряда, где все синонимы, кроме того, что синонимичны доминанте, должны быть синонимичны друг другу, т.е. "линейный" ряд должен выстраиваться "вкруговую". Однако все словари синонимов построены по принципу "линейности" ряда и начинаются с доминанты.

Если строить словарь по доминанте, то встает вопрос о формальных признаках ее выделения в толковых словарях. Логически выделяется два таких признака:

1) не иметь в толковании "семы", совпадающей с "ядерной семой" доминанты;

2) не иметь помет (хотя такие случаи все-таки встречаются [5]).

Если не ставить себе задачей "кругового" построения синонимического ряда, то удобно будет использовать в качестве доминант лексемы семантических полей, например, "Части тела человека", поскольку эти лексемы наиболее распространены.

2. ***Функциональному*** или ***коннотативному***, т.е. поиск ведется по помете, которая сопровождает толкования слов в академических словарях. Список помет обязательно приводится в любом словаре, их сокращения традиционны.

3. ***Сочетаемостному*** или ***синтаксическому***. Последний признак очень сложно формализовать, поскольку, хотя сведения о сочетаемости слов и подаются в толковых словарях (напр., об управлении глаголов), они не имеют одинакового стандартного оформления и зачастую нерегулярны.

Сравнивая выделенные нами принципы построения синонимического ряда из сло́вника толкового словаря с требованиями, предъявляемыми к лексическим единицам, трактуемым как синонимы в Новом объяснительном словаре синонимов русского языка, приведенными выше в определении, а также следующими [5]:

а) совпадающая часть толкований всех лексем ряда должна относиться к их ассерции (та часть значения лексемы, которая подвергается отрицанию);

б) она должна включать родовое понятие;

в) в случае, если имя родового понятия синтаксически зависит от какого-то другого слова толкования, требуется, чтобы это синтаксически главное слово тоже совпадало.

Таким образом, можно прийти к выводу, что на сегодняшний день невозможно полностью формализовать процесс построения синонимического ряда. Мы, тем не менее, предпримем такую попытку, сравнив наш результат с соответствующими словарными статьями синонимических словарей. Работа предлагаемой программы, основанной на изложенных выше принципах, схематически может быть представлена следующим образом:

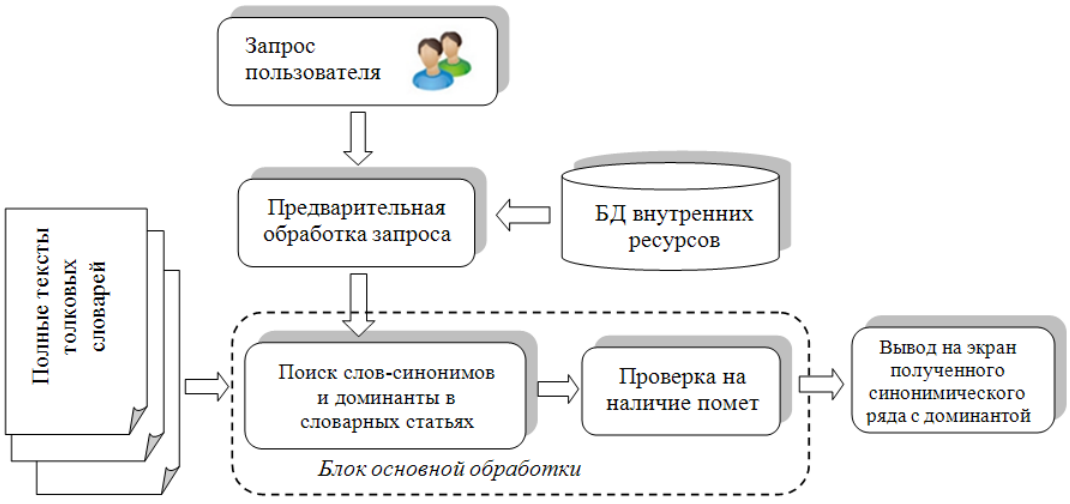

Рисунок 5 − Схема работы программы, выделяющей синонимические ряды из толковых словарей

Согласно алгоритму, работа программы состоит в следующем. Толковый словарь загружается программой в момент ее запуска. Пользователю предоставляется возможность выбрать слово из списка (1) (рис. 6), который представляет собой перечень слов из загруженного словаря.

При нажатии на слове из списка в окне "Словарная статья" пользователь сможет увидеть словарную статью, относящуюся к данному слову (2). Если ему необходимо узнать синонимы для этого слова, то нужно выбрать слово и нажать на кнопку "Синонимический ряд". Тогда в окне (3) появится список синонимов (синонимический ряд), который соответствует выбранному слову.

Также в меню программы можно узнать, как пользоваться программой, теоретическую информацию о синонимах и загрузить дополнительные словари в текстовом формате.

Окно программы представлено на рис. 6.

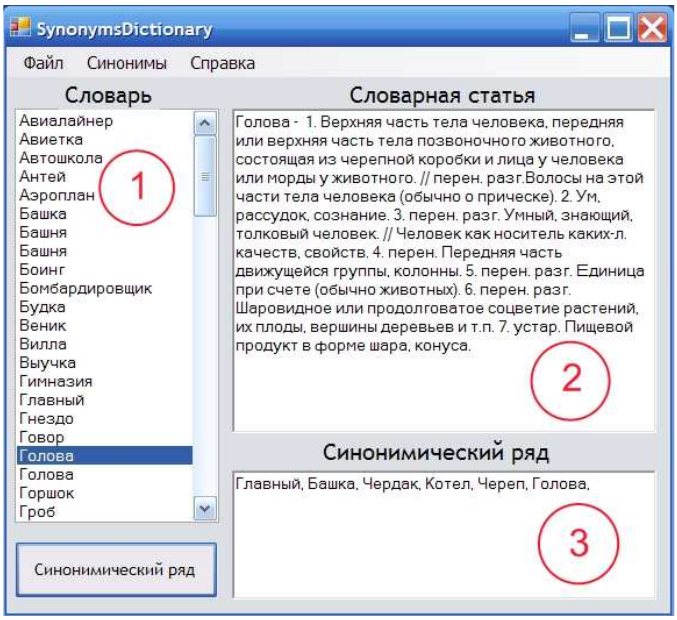

Рисунок 6 – Интерфейс программы атоматической генерации синонимического ряда

Основные результаты и выводы. Проведенное исследование показало, что автоматизированные методы создания словарей применимы для создания словарей синонимов. Авторами предложен подход к автоматическому формирования синонимического ряда, с использованием структуры словарной статьи толкового словаря. Предложенный метод был программно реализован с помощью языка программирования C#.

ЛИТЕРАТУРА:

1. Александрова З. Е. Словарь синонимов русского языка / Под. ред. Л. А. Чешко. − Изд. 4-е, репродуцированное. − М. : "Русский язык", 1975. − 600 с.
2. Алефиренко Н. Ф. Теория языка / Н. Ф. Алефиренко – М. : Академия, 2004. – 368 с.
3. Словник синонімів української мови: В 2 т. / А. А. Бурячок, Г. М. Гнатюк, С. І. Голова щук та інші. − К.: Наук. Думка, 2001. − Т. 1. − 1040 с.; Т. 2. − 960 с.
4. Селіванова О. Сучасна лінгвістика: термінологічна енциклопедія. – Полтава : Довкілля-К, 2006. − 716 с.
5. Новый объяснительный словарь синонимов русского языка. Второе издание, исправленное и дополненное / Под общим руководством акад. Ю. Д. Апресяна. − М. : Школа "Языки славянской культуры", 2003. − 1488 с.
6. Стернин И. А. Лексическое значение слова в речи. − Воронеж, 1985. − С. 36-85.
7. Словарь ABBY Lingvo 12. Электронный ресурс : http://www.lingvo.ua/ru/LingvoDictionaries/Details?dictionary=LingvoUniversal%20%28En-Ru%29
8. Тришин В. Н. Электронный словарь-справочник синонимов русского языка системы ASIS. Версия 6.0 (последнее обновление 13.11.2010). / В. Н. Тришин − Москва, 1993−2010. Электронный ресурс : http://www.trishin.ru

БОРИСОВА Наталья Владимировна – ассистент кафедры интеллектуальных компьютерных систем Национального технического университета "Харьковский политехнический институт".

Научные интересы: компьютерная лексикография, искусственный интеллект, автоматизация библиотечной деятельности.

КАНИЩЕВА Ольга Валерьевна – к.т.н., доцент кафедры интеллектуальных компьютерных систем Национального технического университета "Харьковский политехнический институт".

Научные интересы: автоматизированная обработка естественного языка, математическое моделирование в лингвистике, искусственный интеллект.

ЮРЧЕНКО Елена Николаевна – к.филос.н., доцент кафедры интеллектуальных компьютерных систем Национального технического университета "Харьковский политехнический институт".

Научные интересы: формальное представление объектов естественного языка, когнитивная лингвистика, знаково-символьные системы естественного языка.